Mentis-Verlag, 2006. S. 171-186.

Texttechnologie an der Universität Bielefeld

Abstract

In the mid-1990s, the Faculty of Linguistics and Literary-Studies at Bielefeld University began to establish the field *Text technology*, both in research and education. *Text technology* is a new field of research on the border of Computational Linguistics and Computational Philology.
This paper focuses on *Text technology* in academic education. In 2002, *Text Technology* was introduced as a minor subject for B.A. Programs. It is organized in modules: Module 1 introduces the characteristics of electronic texts and documents, typography, typesetting systems and hypertext. Module 2 introduces one or two programming languages relevant to the field of humanities computing. Markup languages and the principles of information structuring are the main topics of Module 3. The formal fundamentals of computer-based text processing, as formal languages and their grammars, Logics et cetera are subjects of another module.
The paper ends with a short description of other Bachelor- and Master-Programs at Bielefeld University which contain text technological themes.

Einleitung

An der Universität Bielefeld ist das Gebiet *Texttechnologie* nicht nur in der Forschung sondern auch in der Lehre vertreten. Der vorliegende Beitrag beschreibt dieses Gebiet insbesondere aus der Perspektive seiner curricularen Umsetzung.

Vom Wintersemester 1999/2000 bis zum Sommersemester 2002 konnte an der Universität Bielefeld das Fach *Texttechnologie* als eines von zwei Nebenfächern eines Magisterstudiengangs studiert werden. 2002 wurde die Universität Bielefeld als eine von zwei nordrhein-westfälischen Modelluniversitäten zur Umsetzung einer neuen Lehramtsausbildung ausgewählt. Hierfür wurden die alten Lehramtsstudiengänge abgeschafft und durch konsekutive Studiengänge ersetzt. Zukünftige Lehrerinnen und Lehrer studieren seither in einem ersten Ausbildungsschritt einen nicht auf die Lehrerausbildung spezialisierten Bachelorstudiengang, der aus einem Kernfach und einem Nebenfach besteht. Sowohl das Kernfach als auch das Nebenfach müssen Schulfächer (zum Beispiel Mathematik, Germanistik) oder Pädagogik sein. Im Zuge dieses Modellversuchs wurden 2002 auch alle Magisterstudiengänge abgeschafft. Seither werden Schritt für Schritt auch die bestehenden Diplomstudiengänge eingestellt. Die Mehrzahl der eingestellten Studienfächer wird heute in konsekutiven Studiengängen weiter angeboten. Im Gegensatz zu der Abschaffung der alten Lehramts- und Magisterstudiengänge war die Einrichtung der neuen Studiengänge allerdings nicht auf einen Automatismus des Modellprojektes zurückzuführen. Alle neuen Studiengänge wurden neu konzipiert und installiert. Im Verlauf dieses Prozesses wurden Inhalte des alten Magisterstudiengangs *Texttechnologie* in verschiedenen neuen Studiengängen aufgenommen.

Der Artikel gliedert sich wie folgt: Nach einer Einführung des Begriffes ›Texttechnologie‹ wird das mit diesem Terminus bezeichnete Gebiet dargestellt. Anschließend werden das ehemalige Magisternebenfach und das gegenwärtig studierbare Bachelornebenfach skizziert und deren Inhalte detailliert aufgeführt. Die weiteren Studiengänge, die texttechnologische Inhalte vermitteln, werden am Ende dieses Beitrages kurz angesprochen.

Das Gebiet *Texttechnologie* an der Universität Bielefeld

Texttechnologie bezeichnet einen Forschungsbereich an der Schnittstelle von Computerlinguistik und Computerphilologie, wobei der Überlappungsbereich zur Linguistik größer ist als zu den Philologien. Wie der Name des Gebietes schon erwarten lässt, steht die maschinelle Verarbeitung geschriebener Sprache im Zentrum dieser noch sehr jungen Disziplin. Für eine effiziente und auch auf Texttypen abgestimmte adäquate Verarbeitung von Texten ist es derzeitig gängige Praxis, diese durch die Hinzufügung von Zusatzinformationen aufzubereiten. Hierbei spielt die Verwendung von Auszeichnungssprachen eine zentrale Rolle.

An der Universität Bielefeld ist das Gebiet *Texttechnologie* an der Fakultät für Linguistik und Literaturwissenschaft angesiedelt. Diese Fakultät ist die größte der fünfzehn an der Universität Bielefeld vorhandenen Fakultäten. Derzeit sind die Fächer Germanistik, Anglistik, Deutsch als Fremdsprache, Literaturwissenschaft und Linguistik an dieser Fakultät vertreten. In den vergangenen Jahren wurden die Fächer Latein, Slawistik und Romanistik abgeschafft, Inhalte dieser Fächer sind in anderen Fächern enthalten und sollen auch erhalten bleiben.

Das Fach Linguistik unterteilt sich in die Arbeitsbereiche *Semantik und Syntax*, *Kommunikation*, *Psycholinguistik und Klinische Linguistik* und *Computerlinguistik und Texttechnologie*. In dem letztgenannten Bereich, der von Dieter Metzing geleitet wird, bündeln sich die Forschungs- und Lehraktivitäten mit Texttechnologiebezug. Neben der Professur sind bis zu drei Mitarbeiter/innen in dem Bereich tätig. Zum Sommersemester 2004 wurde eine Juniorprofessur eingerichtet, die mit Alexander Mehler besetzt wurde.

Forschung

Im Zentrum der Forschungsaktivitäten des Bereichs *Computerlinguistik und Texttechnologie* steht die verteilte Forschergruppe *Texttechnologische Informationsmodellierung*. Dieses seit dem Jahr 2001 von der *Deutschen Forschungsgemeinschaft* (DFG) finanzierte Verbundprojekt ist dezentral strukturiert, mit Projekten in Tübingen/Osnabrück, Gießen und Dortmund. Der Kern der Forschergruppe ist in Bielefeld, wo sich zwei Projekte befinden. Die Forschergruppe wird von Dieter Metzing geleitet. Im Jahr 2005 wurde die Forschergruppe durch die DFG evaluiert und die Finanzierung bis 2008 zugesichert. Um diese Forschergruppe herum sind weitere Forschungsaktivitäten angesiedelt.[1]

Entwicklung der Texttechnologiestudiengänge

Der Bachelorstudiengang *Texttechnologie* im Nebenfach existiert seit dem Wintersemester 2002/2003. Er basiert auf dem zum Wintersemester 1999/2000 eingeführten Magisternebenfachstudiengang *Texttechnologie*, dem letzten an der Universität eingeführten Magisterstudiengang. Das B.A.-Nebenfach *Texttechnologie* wurde wie schon vorher der Magisternebenfachstudiengang so konzipiert, dass Theorie und Berufsorientierung miteinander verbunden sind. Das schlägt sich in den Veranstaltungen des Studiengangs, ihrer Thematik und in ihren Veranstaltungsformen nieder. Der Studiengang wurde von Anfang an – auch außerhalb Bielefelds – als äußerst interessanter und innovativer Studiengang wahrgenommen. Vorgestellt wurde er unter anderem auf der gemeinsamen Jahrestagung der *Association of Literary and Linguistic Computing* (ALLC) und der *Association for Computing in the Humanities* (ACH), die 2005 in Victoria (Kanada) stattfand.[2]

Am 12. und 13. Dezember 2005 erfolgte eine Begehung der Fakultät durch die Akkreditierungsorganisation ZEvA und von ihr beauftragte externe Experten. Zum Zeitpunkt der Abfassung dieses Beitrags liegt jedoch der schriftliche Bericht der Akkreditierungsorganisation noch nicht vor. Die Lehrenden des Studiengangs erwarten aber, dass der Studiengang *Texttechnologie* positiv bewertet und akkreditiert werden wird. Am Rande der Akkreditierung wurde positiv angemerkt, dass die B.A.-Studiengänge an der Bielefelder Fakultät für Linguistik und Literaturwissenschaft in vorbildlicher Weise Kontinuität und Wandel verbunden hätten. Für den Texttechnologiestudiengang bedeutet Kontinuität und Wandel, dass einerseits die Studieninhalte des Magisternebenfachs in möglichst umfassender Weise in das B.A.-Nebenfach integriert wurden, allerdings andererseits das neue, in Module strukturierte und mit Leistungspunkten bewertete B.A.-System übernommen wurde. Dies mag den Vorwurf provozieren, es handele sich um »alten Wein in neuen Schläuchen« (wobei im Falle des Magisternebenfachs *Texttechnologie* der älteste Jahrgang *1999* ist). Jedoch bietet die neue Struktur inhaltliche Zusammenfassungen, die das Studium übersichtlicher werden lassen. Es werden damit von Beginn an Zusammenhänge von Lehrveranstaltungen sichtbar, die im Magisterstudium nur implizit erkennbar waren.

Das Magisternebenfach *Texttechnologie*

Das Magisternebenfach *Texttechnologie* gliederte sich, wie alle Magisterstudiengänge, in zwei Teile, das Grund- und das Hauptstudium. Im Grundstudium wurden die Veranstaltungen *Grundkurs Texttechnologie*, *Formale Methoden I & II*, *Hypertext* und *Textgestaltung und Textsatz* studiert. Das Grundstudium wurde mit einer Prüfung abgeschlossen.

Das Hauptstudium begann mit den Seminaren *Basisveranstaltung Texttechnologie I: Methoden und grundlegende Standards* und *Basisveranstaltung Texttechnologie II: Weiterführende Standards*. Parallel dazu oder auch im Anschluss an die Basisveranstaltungen konnte die *Einführung in die Programmierung* besucht werden. Den Abschluss des Nebenfachstudiums bildete ein Projektseminar. Mit der Ausnahme des das Magisternebenfach abschließenden Projektseminars finden sich alle Veranstaltungen auch im Bachelor-Nebenfach und werden ausführlich im nächsten Abschnitt beschrieben.

Eine dem Magisterstudium eigene Besonderheit, die durch die Umstellung auf den Bachelorstudiengang verloren gegangen ist, war die Möglichkeit, computerphilologische Themen in einer Magisterarbeit umfassend zu behandeln. Viele Studierende schrieben in ihren Hauptfächern (zum Beispiel Literaturwissenschaft, Anglistik, Germanistik oder Linguistik) eine Magisterarbeit, integrierten jedoch eine Vielzahl von texttechnologischen Inhalten oder verwendeten Methoden und Techniken, die Bestandteil des Nebenfachstudiengangs *Texttechnologie* waren. Beispielhaft seien hier nur einige Titel derartiger Magisterarbeiten genannt: im Hauptfach Germanistik wurden Arbeiten zu *Mündlichkeit und Schriftlichkeit in computervermittelter Kommunikation* und zu *Tageszeitungen im Internet*, in Literaturwissenschaft zu *Literatur - Computer - Architektur: Neue Ästhetik für Neue Medien* und zum *Mehrwert von Metadaten im Kontext der Drehbuchentwicklung* und im Hauptfach Linguistik zum *Elektronischen Publizieren von Wörterbüchern* verfasst und von Lehrenden aus dem Bereich *Texttechnologie* (mit-)betreut.

Das nachfolgend ausführlich dargestellte Bachelornebenfach erlaubt im Prinzip ebenfalls die schriftliche Ausarbeitung computerphilologischer Inhalte, da eine umfangreichere Hausarbeit integralerer Bestandteil des B.A.-Kernfachs ist. Diese Arbeit hat allerdings einen wesentlich geringeren Umfang als eine Magisterarbeit. Darüber hinaus ist die Bache-lorarbeit eng mit

einer Lehrveranstaltung des Kernfachs verwoben. Aus diesen Gründen hat es bisher keine B.A.-Arbeiten gegeben, die neben den Inhalten des Kernfachs auch im größeren Umfang Inhalte des Nebenfachs *Texttechnologie* thematisierten.

Das Bachelor-Nebenfach *Texttechnologie*

Die Konzeptionierung des Faches *Texttechnologie* als Nebenfach eines Bachelor-Studiums mit bevorzugt geisteswissenschaftlichem Kernfach zeigt die ergänzende Funktion des Studienfaches zu philologischen Studienschwerpunkten auf.

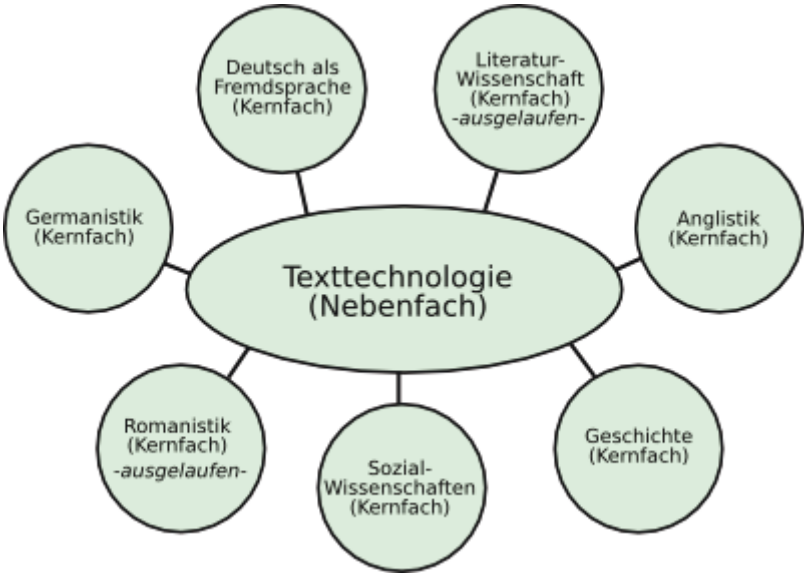

Abbildung 1: Gewählte Fächerkombinationen von Studierenden des Nebenfachs Texttechnologie

Abbildung 1 zeigt schematisch die von den Studierenden des Bachelornebenfachs gewählten Studienfachkombinationen. Am häufigsten wird *Texttechnologie* als Nebenfach zu den Hauptfächern Germanistik, Anglistik und Literaturwissenschaft gewählt, wobei das letztgenannte Fach seit 2005 nicht mehr als Bachelorstudiengang angeboten wird, sondern ausschließlich als M.A.-Studiengang. Das in der Abbildung ebenfalls als *ausgelaufen* markierte Fach Romanistik ist im Jahr 2004 vollständig abgeschafft worden. Damit setzte die Universität Bielefeld eine Entwicklung fort, die mit der Abschaffung des Faches Slawistik begann und ihre erste Fortsetzung in der Abschaffung des Faches Latein fand. Diese Entwicklung ist von verschiedenen Seiten bedauert worden.

Das B.A.-Nebenfach *Texttechnologie* dient in erster Linie der Erweiterung philologischen Wissens um Kenntnisse über Verfahren der rechnergestützten Verarbeitung textbasierter Informationen, versucht dabei jedoch über den Status einer bloßen Zusatzqualifikation hinaus das Verständnis von Texten im Kernfach grundlegend zu erweitern und dem Studierenden neue Sichtweisen auf Texte zu ermöglichen.

Die Entscheidung für das Nebenfach *Texttechnologie* bedeutet in Hinblick auf die Wahl eines philologischen Kernfaches meist eine Berufsorientierung in Richtung publizistischer Tätigkeiten; einer der zwei philologischen Berufsfeldschwerpunkte neben der Wissensvermittlung an Schulen und Hochschulen.

Die meist geisteswissenschaftliche Orientierung der Studierenden, die nicht zwangsläufig eine informatische Vorbildung mitbringen, erfordert eine sanfte Heranführung an die, über normale EDV-Tätigkeiten hinausgehende Arbeit mit dem Computer. So werden keine umfangreichen Computerkenntnisse für das Studium vorausgesetzt, wenngleich eine Affinität für das Medium Computer ein klarer Vorteil ist.

Das Bachelor-Nebenfach *Texttechnologie* in Bielefeld ist auf eine Studienzeit von 6 Semestern ausgerichtet, mit einem empfohlenen Studienstart im Wintersemester. Den internationalen Standards des Bachelor-Systems folgend, gibt es eine Leistungsprotokollierung nach dem *European Credit Transfer System* und studienbegleitende Lernkontrollen, die in Form von Einzelleistungsnachweisen durch Hausarbeiten, Referate, Projekte, Klausuren oder Übungen erbracht werden können.

Das Studium ist in Module aufgegliedert, die Kurse zu einem bestimmten Themengebiet zusammenfassen. In vier Basismodulen werden die Grundkenntnisse des texttechnologischen Arbeitens vermittelt. In einem Wahlpflichtmodul können sich die Studierenden für einen Vertiefungsschwerpunkt ihres Nebenfachstudiums entscheiden.

Zum Teil sind diese Module in vorgegebener Reihenfolge zu studieren, da einige Module den erfolgreichen Abschluss anderer voraussetzen. So ergibt sich ein für die meisten Studierenden im Nebenfach Texttechnologie prototypischer Studienverlauf über 6 Semester (siehe Abbildung 2).

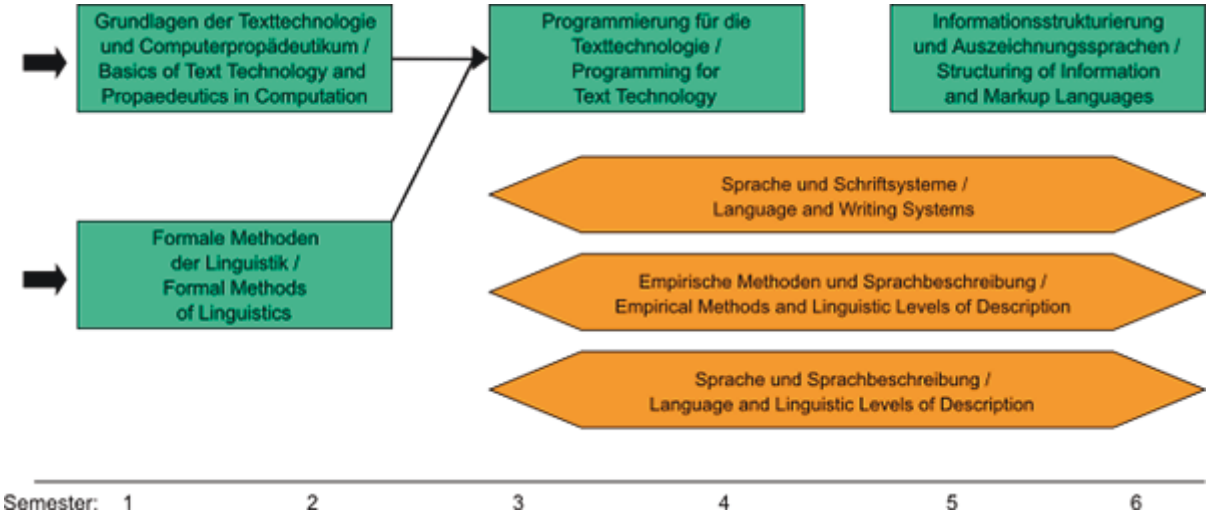

Abbildung 2: Schematische Darstellung des Ablaufs des Nebenfachstudiengangs

Den Einstieg in die Texttechnologie schafft das Modul TT1: *Grundlagen der Texttechnologie und Computerpropädeutikum*. Es besteht aus drei Kursen: Im ersten Semester ist vorgesehen im Rahmen eines *Computerpropädeutikums* grundlegendes Wissen über Computersysteme und den Umgang mit ihnen zu vermitteln. Neben Kenntnissen über Betriebs- und Dateisysteme, die viele der Studierenden bereits mitbringen, werden auch texttechnologische Grundlagen gelehrt, etwa Wissen um Zeichensatzsysteme (ASCII, Unicode) und der Umgang mit komplexen, textverarbeitenden Systemen.

Zur Übung regulärer Ausdrücke und als Beispiel eines komplexen, textverarbeitenden Systems gibt es eine Einführung in *Emacs*, einen gerade in der Linux-Welt populären,

kostenfreien Editor mit umfangreichen Möglichkeiten zur Textmanipulation. Die versierte Handhabung dieses Tools wird in späteren Veranstaltungen zu Auszeichnungssprachen oder zur Programmierung vorausgesetzt, darauf aufbauend werden weitere, spezifische Editor-Systeme je nach Anwendungsbereich eingeführt (etwa *Oxygen* für die Verarbeitung von XML/XSL oder *Eclipse* als Entwicklungsumgebung für objekt-orientierte Programmierung).

Begleitend hierzu werden in einer *Einführung in die Texttechnologie* Eigenschaften elektronischer Dokumente und Methoden zu deren Erstellung und Verarbeitung dargelegt. Außerdem wird allgemein das Gebiet der *Texttechnologie* vorgestellt. So bekommen die Studierenden gleich zu Beginn ihres Studiums eine Übersicht über die Aufgabenvielfalt ihres Studiums und einen Ausblick auf Themen zukünftiger Lehrveranstaltungen.

Im dritten Kurs des Moduls, *Hypertext*, das im folgenden, zweiten Semester studiert werden sollte, werden Begriffe, Konzepte, Definitionen und die Geschichte hypertextueller und hypermedialer Systeme (*Memex*, *Xanadu*, *Hypercard* ...) bis hin zum *World Wide Web* gelehrt. Die Nutzung hypermedialer Strukturen wird praktisch am Beispiel der *HyperText Markup Language* (HTML) unterrichtet. Dazu gehört eine Einführung in externe Layoutdefinition zur Trennung von Inhalt und Form (am Beispiel der verbreiteten *Cascading StyleSheets*, CSS) ebenso wie die Vorstellung erweiterter, clientseitiger Interaktionsmöglichkeiten (unter anderem am Beispiel von *JavaScript*).

Textstruktur und Textsatz führt im zweiten Semester tiefer in das Dokumentkonzept hinein. Neben Grundlagen klassischer Typografie, modernem Layout und deren Fachterminologien, werden Techniken zur Einhaltung des Paradigmas der Trennung von Inhalt und Design anhand verschiedener Dokument-Systeme (*Microsoft Word*, *OpenOffice*, *LaTeX*, *PostScript*, *PDF*) vorgestellt und geübt. Besonders die Verwendung von *LaTeX* als verbreitete Auszeichnungssprache jenseits der SGML-Sprachfamilie wird geübt und daran der Unterschied zwischen visuellem und generischem Markup gelernt.

Die *Einführung in die Texttechnologie* wird mit einer Klausur abgeschlossen. Im zweiten Semester können die Studierenden wählen, ob sie entweder in *Hypertext* oder *Textstruktur und Textsatz* eine abschließende mündliche Prüfung ablegen.

Neben dem gelernten Wissen um die Grundlagen des texttechnologischen Arbeitens sind die Studierenden nun in der Lage, eine eigene hypertextuell-struktierte Internet-Präsenz zu erstellen um eigene Erfahrungen mit Web-Publishing zu machen. Hierfür bietet die Universität Bielefeld allen Studierenden Speicherplatz auf ihrem Webserver an, der frei durch clientseitig verarbeitete Inhalte (in der klassischen Kombination HTML+CSS+*JavaScript*), aber auch vorausblickend auf serverseitige Programmierung in späteren Kursen (PHP, Perl) genutzt werden kann.

Parallel zu dem praktischen Modul TT1 werden in den ersten beiden Semestern auch stark theoretische Inhalte im Modul Lin2: *Formale Methoden der Linguistik* vermittelt.

Im ersten Kurs, der begleitend zur *Einführung in die Texttechnologie* studiert wird, gibt es zunächst eine Einführung in die Mengentheorie - oftmals eine Wiederholung bereits in der Mittel- oder Oberstufe gelernter Grundlagen in Verbindung mit nötigen Formalismen und der Fachterminologie. Themen dabei sind analytische Strategien einer einfachen Logik, Mengenlehre mit Mengenoperationen, mengentheoretischen Gesetzen, Relationen und Funktionen, Partitionen und Äquivalenzklassen. Ein weiterer Schwerpunkt ist eine Einführung in die Theorie formaler Sprachen. Der Chomsky-Hierarchie formaler Sprachen

folgend wird mit Beschreibungsarten von Grammatiken regulärer Sprachen begonnen, etwa den auf Kleene-Operatoren basierenden regulären Ausdrücken (Regular Expressions), die im Rahmen des *Computerpropädeutikums* begleitend praktisch trainiert werden, endlichen Übergangsnetzwerken (Finite State Networks/FSN), endlichen Automaten (Finite State Automata/FSA) und regulären Grammatiken (lineare Grammatiken, Typ-3-Grammatiken). Weiterführend werden endliche Maschinen (Finite State Transducer/FST) und kontextfreie Grammatiken (zur Formulierung von Typ-2-Sprachen der Chomsky-Hierarchie) vorgestellt.

Im zweiten Kurs des Moduls, der üblicherweise im zweiten Semester studiert wird und Wissen aus dem ersten Kurs voraussetzt, wird verstärkt auf die Formalismen der Logik eingegangen. Grundlagen der Aussagen- und Prädikatenlogik mit ihren Gesetzen, Operatoren und Interpretationsfunktionen werden vermittelt, Fachtermini wie ›Konklusion‹, ›Tautologie‹ und ›Kontradiktion‹ erläutert. Anhand von Dialogspielen und anhand des Kalküls des natürlichen Schließens werden die gelernten Formalismen in komplexeren Zusammenhängen geübt.

Der dritte Kurs des Moduls *Formale Methoden* wird parallel zum zweiten Kurs im zweiten Semester studiert. Behandelt wird der Aufbau von Lexika mittels komplexer Attribut-Wert-Strukturen (AWM). Konzepte wie Identität, Generalisierung und Subsumption werden erklärt, ebenso Funktionen über AWMs wie Unifikation, Differenz oder Typ-Vererbung. Die Konstruktion gerichteter azyklischer Graphen wird als typische Visualisierung von AWMs gelehrt. Anhand verschiedener, auf Attribut-Wert-Strukturen basierender Grammatiktheorien wie der *Head-Driven Phrase Structure Grammar* (HPSG) oder der *Lexical Functional Grammar* (LFG) werden die Funktionen von AWMs anschaulich geübt.

Gerade geisteswissenschaftlich orientierten Studierenden fällt die mathematisch-logische Betrachtung von Sprache, die im Rahmen der formalen Methoden vorgestellt wird, oft schwer, weshalb diese intensiv durch Aufgabenblätter und in Tutorien geübt werden, um eine breite Basis für weiterführende texttechnologische Studien bilden zu können.

Jeder Einzelkurs wird mit einer Klausur abgeschlossen.

Nach der Vermittlung theoretischen Wissens wird im Modul TT2: *Programmierung für die Texttechnologie* praktisch die Verarbeitung strukturierter, annotierter Textdaten geprobt. Dazu ist ein Einführungskurs in eine Programmiersprache zu besuchen, der auch Studierenden ohne Programmiererfahrung umfassende Kenntnisse zum Erstellen funktionierender Programme und Basiswissen zur Implementation textverarbeitungsrelevanter Algorithmen vermittelt. Im darauffolgenden Semester ist ein umfangreiches Programm zu entwickeln und zu dokumentieren, das gelernte Verfahren zur Textdatenverarbeitung implementiert.

Die Wahl der Programmiersprache steht den Studierenden offen – gängige Sprachen im Studienangebot des Fachbereichs *Texttechnologie* sind *Java*, *Perl*, PHP, *Prolog* oder XSL-T.

Jede Sprache ist dabei für unterschiedliche Zwecke optimiert und in unterschiedlichem Grad spezialisiert. *Java* dient, als Beispiel einer objektorientierten Sprache, der Programmierung komplexer Desktop-Systeme. *Perl* und PHP, als nativ imperative Programmiersprachen, zeigen ihre Stärken in Webanwendungen mit Schnittstellen zu den bereits bekannten Sprachen HTML, *JavaScript* und CSS, aber auch, besonders im Fall *Perl*, bei umfangreichen Stringmanipulationen. Dadurch wurde diese Sprache gerade im Bereich der *Texttechnologie* sehr populär.[3] *Prolog*, als deklarativ-logische Sprache, bietet sich für direkte Implementationen der in den formalen Methoden der Linguistik erlernten Formalismen (FSN,

FST, AWM) hervorragend an. XSL-T letztlich, dem deklarativ-funktionalen Paradigma folgend, ist auf die Transformation XML-annotierter Dokumente in andere Zieldokumente spezialisiert und so gerade in der Korpusbearbeitung von großem Nutzen.

Neben diesen eigenständigen Programmiersprachen wird auch die Modellierung und Nutzung von Datenbanken mithilfe der Abfragesprache SQL gelehrt.

Das Modul TT3: *Informationsstrukturierung und Auszeichnungssprachen* fasst zwei Kurse zusammen, in denen Methoden zur Modellierung und Bearbeitung informationell angereicherter Texte vorgestellt werden.

Nach dem Musterstudienplan im fünften, aber üblicherweise bereits im dritten Semester werden standardisierte *Auszeichnungssprachen* studiert. Darin werden die heute relevanten Standards *Standard Generalized Markup Language* (SGML) und *eXtensible Markup Language* (XML) ebenso wie Varianten ihrer Grammatik (DTD, XML Schema, RelaxNG) und allgemeiner Konzepte (Wohlgeformtheit, Gültigkeit, Namensräume) vorgestellt. Schnittstellen für den Dokumentenzugriff (DOM, SAX, XPath), Möglichkeiten ihrer Transformation (über XSL) und erweiterte Spezifikationen (XLink) werden praktisch geübt.

Auf Basis dieser Kenntnisse werden im darauf folgenden Semester in einer vierstündigen Veranstaltung *Informationsstrukturierung* konkrete Anwendungsbereiche von Auszeichnungssprachen besprochen.

Konzepte zur dokument-, daten- und objektorientierten Modellierung werden diskutiert und Nutzbarkeit in automatischen Textanalyse-Systemen zur Informationsextraktion, Exploration oder zum Information-Retrieval erläutert. Die Genese von Strukturmodellen in drei Schritten, von einem konzeptionellen, über ein logisches hin zu einem physikalischen Modell, soll in Gruppen geübt und verstanden werden, wobei die *Rhetorical Structure Theory* (RST) als grammatisches Modellbeispiel zugrundeliegen kann oder die Arbeit der *Text Encoding Initiative* (TEI) als Beispiel komplexer Dokumentmodellierung.

Zur Visualisierung strukturierter Informationen wird die *Unified Modelling Language* (UML) eingeführt.

Als Beispielsprache für komplexe, dokumentorientierte Modellierung wird das Wissen aus Auszeichnungssprachen in Bezug auf XML Schema vertieft.

Das Modul wird mit einer Klausur im Kurs *Auszeichnungssprachen* abgeschlossen. In *Informationsstrukturierung* ist eine Projektarbeit vorzustellen, die im Rahmen der Gruppen erarbeitet wurde.

Abbildung 3: Seminarraum mit 40 Rechnerarbeitsplätzen für Studierende

Die Mehrzahl der praktische und theoretische Inhalte vermittelnden Seminare aus dem Texttechnologiestudium, zu denen insbesondere die Seminare aus den Modulen TT1, TT2 und TT3 gehören, finden in speziellen Seminarräumen statt, die es den Lehrenden erlauben, praktische Inhalte direkt am Rechner vorzuführen und die es den Studierenden ermöglichen, praktische Komponenten selbst auszuprobieren. (siehe Abb. 3) Durch den Einsatz freier Software oder den Erwerb von speziellen Lizenzen wird dafür gesorgt, dass die im Studium verwendete Software auch für Übungen zu Hause genutzt werden kann. Gerade hierfür ist es sinnvoll, die relevanten Programme kurz testen zu können.

Üblicherweise im vierten Semester, unter Umständen jedoch schon früher, können sich die Studierenden eine Schwerpunktsetzung in ihrem Nebenfachstudium wählen. Grundsätzlich ist diese Wahl offen, wird aber nicht selten von den Anforderungen der Bachelorarbeit im Kernfach beeinflusst. Sind dort etwa umfangreiche Korporaverarbeitungen von Nöten, liegt eine Profilierung durch das Modul Lin4.1: *Empirische Methoden und Sprachbeschreibung* nahe. Ist die Verarbeitung mehrsprachiger Daten (vielleicht in verschiedenen Schriftsystemen vorliegend) und deren Vergleich Bestandteil der Arbeit, bietet das Modul Li1: *Sprachen und Schriftsysteme* die Möglichkeit, sich näher mit den behandelten Sprachen und Schriftsystemen zu befassen. Sind hingegen makro- und mikrostrukturelle Betrachtungen von Texten aus linguistischer Sicht interessant, können in Modul Lin3.1: *Sprache und Sprachbeschreibung* hierfür fortgeschrittene Werkzeuge kennengelernt werden.

Das Modul *Empirische Methoden und Sprachbeschreibung* vermittelt Wissen um Methoden und Werkzeuge der Empirie, die für die Evaluierung und die statistische Auswertung annotierter Korpora von großer Bedeutung sind. Des Weiteren werden in diesem Modul

Techniken der deskriptiven Linguistik zur Sprachbeschreibung auf verschiedenen linguistischen Betrachtungsebenen vorgestellt. Neben einem zweigeteilten Kurs *Empirische Methoden der Linguistik* ist ein weiterer Kurs aus dem Bereich *Syntax und Morphologie*, *Phonetik und Phonologie* oder *Semantik und Pragmatik* für dieses Modul zu studieren.

Aufbauend auf den Kenntnissen der formalen Methoden werden im Wahlpflichtmodul *Sprache und Sprachbeschreibung* Techniken der Sprachbetrachtung und -beschreibung auf verschiedenen linguistischen Ebenen gelehrt, zu denen die *Phonetik* und die *Phonologie*, die *Syntax* und die *Morphologie*, sowie die *Semantik* und die *Pragmatik* zählen. Diese Informationen dienen dem differenzierteren Anreichern von Texten und der komplexeren Auswertung von auf diesen Ebenen annotierten Dokumenten. Neben zwei Kursen aus diesem Bereich ist ein Sprachkurs zu besuchen, so dass die gelernten Techniken auf eine zu lernende Sprache appliziert werden können.

Im Profilmodul *Sprachen und Schriftsysteme* können umfangreiche Kenntnisse über fremde Sprachen mit zum Teil nicht-lateinischen Schriftsystemen erworben werden, die notwendig sind, um multilingual annotierte Korpora verwerten zu können. Gerade bei der rechnergestützten Umsetzung nicht auf dem lateinischen Schriftsystem basierender Zeichensysteme fallen Grenzen von im lateinischen Schriftraum üblichen Annotations-Standards auf. Doch auch andere morphologische oder syntaktische Sprachphänomene wollen bedacht sein. Zu diesem Zweck werden in diesem Modul verschiedene Fremdsprachen angeboten. An der Universität Bielefeld werden regelmäßig Kurse in Russisch, Griechisch, Arabisch oder Japanisch angeboten. Mindestens zwei unabhängige Sprachkurse sollten in diesem Modul besucht werden.

Weitere Bielefelder Studiengänge mit texttechnologischen Inhalten

Neben dem Bachelornebenfachstudiengang *Texttechnologie* existieren weitere Studiengänge, die sich mit computerphilologischen und texttechnologische Themen und Methoden auseinandersetzen. Die Abbildung skizziert einen fachlichen Überlappungsbereich *Texttechnologie* dreier auf den ersten Blick sehr unterschiedlicher Studiengänge.

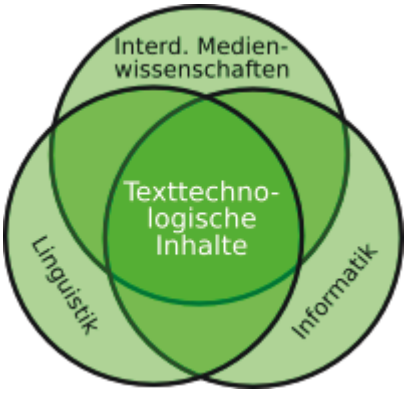

Abbildung 4: Überlappungsbereich *Texttechnologie*

Linguistik ist an der Universität Bielefeld als Bachelor und als Masterstudiengang vorhanden. Im Bachelorstudiengang kann *Texttechnologie* als eines von drei Profilen gewählt werden, ein Studium, das sehr große Gemeinsamkeiten zum Studium des eigenständigen Nebenfachs *Texttechnologie* besitzt. Aufgrund dieser Nähe ist es nicht möglich, das Kernfach *Linguistik* mit dem Nebenfach *Texttechnologie* zu kombinieren. Im Masterstudiengang *Linguistik* kann eine sprachtechnologieorientierte Spezialisierung gewählt werden. Hierbei gibt es inhaltliche

Nähe zur Texttechnologie, jedoch keine Überschneidungen. Im B.Sc.-Studiengang *Kognitive Informatik* kann ein Vertiefungsmodul *Sprachverarbeitung* gewählt werden, im Rahmen dessen auch einige der oben beschriebenen Seminare, insbesondere aus dem Modul *Informationsstrukturierung und Auszeichnungssprachen* studiert werden können.

Die Bezeichnung *Interdisziplinäre Medienwissenschaft* benennt einen M.A./M.Sc.-Studiengang, der bereits erfolgreich akkreditiert wurde. Eine Besonderheit dieses Studiengangs ist es, dass er nicht einer Fakultät zugeordnet sondern fakultätsübergreifend organisiert ist. Je nach Gewichtung der technischen und der geistes- und sozialwissenschaftlichen Studieninhalte wird nach Abschluss des Studiums der Titel *Master of Arts* (M.A.) oder *Master of Science* verliehen. Ein Wahlpflichtmodul dieses Studiengangs trägt die Bezeichnung *Texttechnologien* und beinhaltet sowohl Stoff des B.A.-Nebenfachs *Texttechnologie* als auch sprachtechnologische Inhalte aus dem M.A.-Studiengang *Linguistik: Kommunikation, Kognition und Sprachtechnologie*.

Abschlussbetrachtung

In den vergangenen 10 Jahren wurde an der Bielefelder Fakultät für Linguistik und Literaturwissenschaft im Arbeitsbereich von Dieter Metzing das Gebiet Texttechnologie sehr erfolgreich etabliert. Die von ihm initiierte Forschergruppe wurde 2005 erfolgreich evaluiert und ihre finanzielle Unterstützung durch die DFG bis 2008 verlängert. Ebenfalls im Jahr 2005 erfolgte die Begehung des Texttechnologiestudiengangs durch eine Akkreditierungsorganisation und es steht zu erwarteten, dass hierdurch der Studiengang für die nächsten fünf Jahre auch eine formale Legitimierung erhalten wird. Damit ist – sowohl in der Forschung als auch insbesondere in der Lehre – für die nächsten Jahre eine Perspektive für das Gebiet Texttechnologie an der Universität Bielefeld geschaffen worden. Es kann aber nicht darüber hinweggesehen werden, dass der Studiengang Texttechnologie zu den kleinen Studiengängen, das heißt zu den Studiengängen mit wenigen Studierenden gehört und dass bisherige wie auch bereits angekündigte neue hochschulpolitische Entscheidungen häufig auf die kleineren Bereiche große Auswirkungen haben. Ob diese Auswirkungen für die Texttechnologie positiv oder negativ sein werden oder ob sie ganz ausbleiben, hängt von vielen Faktoren ab. Wichtig wird zweifelsohne sein, dass der Studiengang weiterhin erfolgreich angenommen wird und dass die Absolventen gute Berufsperspektiven haben. Die bisherigen Erfahrungen sind auch in diesem Punkt weitgehend als positiv zu betrachten.

Andreas Witt/Dieter Metzing [4](Bielefeld)

Dr. Andreas Witt
Nauklerstr. 35
Universität Tübingen
72074 Tübingen
andreas.witt@uni-tuebingen.de

Nils Diewald B.A.
Fakultät für Linguistik und Literaturwissenschaft
- Computerlinguistik und Texttechnologie -
Universität Bielefeld
Postfach 10 01 31
33501 Bielefeld
nils.diewald@uni-bielefeld.de

(26. Februar 2006)

Bibliographie

Witt, Andreas/Dieter Metzing:

2005 Texttechnologie in der Universitären Lehre. In: ALLCACH2005, Joint Conference of the ALLC and ACH. Victoria BC.

---

[1] Neben der unter anderem durch die Forschergruppe, aber auch über Verbände wie die *Gesellschaft für linguistische Datenverarbeitung* und deren Arbeitskreis *Texttechnologie*, gegebenen starken nationalen Vernetzung ist die Bielefelder *Texttechnologie* auch in einer vielfältigen Weise international eingebunden. So existieren Forschungskontakte zu Claus Huitfield, dem Leiter der Texttechnologiegruppe der Universität Bergen, zu Allan Renear und David Dubin von der University of Illinois at Urbana-Champaign, zu Michael Sperberg-McQueen (USA) und Felix Sasaki (Japan), die beide in der Architekturgruppe des World Wide Web Consortiums (W3C) tätig sind.

[2] Andreas Witt/Dieter Metzing (2005).

[3] Derzeit ist jedoch in der Texttechnologie eine immer umfassendere Verwendung der Programmiersprache *Python* zu beobachten. Es ist gängige Praxis, dass das Lehrangebot des Studiengangs an derartige Entwicklungen angepasst wird.

[4] Die Autoren dieses Beitrages besitzen unterschiedliche Erfahrungen auf dem Gebiet der Texttechnologie. Andreas Witt war seit 1996 in Bielefeld wissenschaftlicher Mitarbeiter beziehungsweise Assistent und seit 1999 akademischer Studienberater für den Magister- und später für den B.A.-Nebenfachstudiengang *Texttechnologie*. Er wechselte im April 2006 nach Tübingen. Nils Diewald studierte das B.A.-Nebenfach *Texttechnologie* in Kombination mit dem Kernfach Germanistik und gehört zu den ersten Absolventen. Nach Abschluss seines B.A.-Studiums begann er das M.A.-Studium Linguistik: Kommunikation, Kognition und Sprachtechnologie.